Training the Data Scientists of the Future

Ahmet Bulut, Tarık Arıcı, Onur Güzey, Barış Arslan, and Ali Çakmak

Abstract—In this position paper, we share our experience and vision in designing and executing a graduate program specific to Data Science education. The purpose of our efforts is to pave the way in Turkey and across the EMEA region for training the data scientists of the future. For more information, please visit our current program offering at http://ds.sehir.edu.tr

Index Terms—Skills of Data Scientists, Masters in Data Science

1 Introduction

From 2005 to 2010, the digital universe grew from 130 exabytes (EB) to 800 EB. The digital universe will double every two years from now till 2020. In 2020, it will be 40,000 EB, i.e., 40 trillion gigabytes, which is more than 5,200 gigabytes for every man, woman, and child in 2020 [1]. In almost every subsystem currently in use, there is a cyclical process that starts with the (i) acquisition of raw data. It is followed by (ii) processing and transforming of this raw data into information so that we can (iii) drive new insights. With more insights, we are better equipped to (iv) make new and informed decisions. These are the decisions that determine whether customers buy what they need, producers design the right products, city officials deploy the right solutions for bettering urban life, our crops and fields return better yields, and whether we live better. With data playing a central role in advancing civilization, it is appropriate to say that the data has become the new oil. The 4-step cyclical process we described is shaped by Data Science. The idea is to find interesting ways to visualize and present raw data in such a way that enables rapid insight discovery. For example, Figure 1 shows which countries use Celcius scale and which countries use Fahrenheit scale for measuring temperature. This picture is worth more than the text: the majority of the world countries use Celcius. Instead of verbose text and raw data, vivid visuals are more powerful and useful in the process of deriving new and actionable insights.

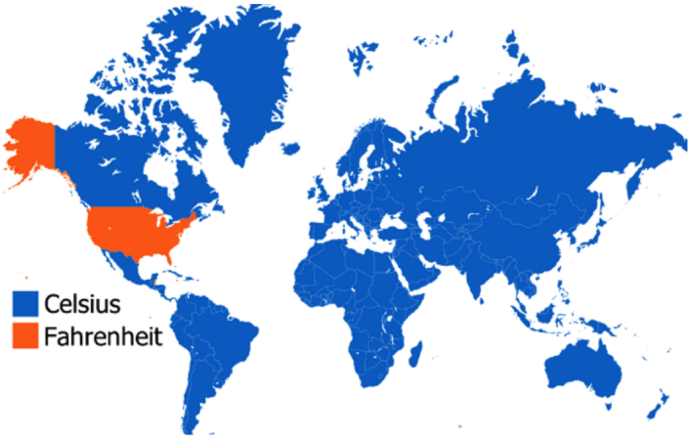


Fig. 1. A picture is worth a thousand words.

2 Skills of a Data Scientist

Data scientists are inquisitive. They explore, ask questions, do what-if analysis, question their existing assumptions and processes. Rather than looking at data from a single source, they examine data from multiple and disparate data sources. All incoming data is sifted through with the goal of discovering hidden insights, which in turn can be used as a competitive business advantage or can provide solutions to pressing business problems. With the necessary analytics and tools support, a capable data scientist will be well equipped to communicate informed conclusions and recommendations across the whole organization [2].

The necessary analytics and tools support varies. From clustering and regression, to classification and probabilistic inference, and to data enrichment and visualisation, data scientists need to have a solid foundation in computer science and applications, modelling, statistics, analytics and mathematics. In order to explore exabytes of data and do what-if analysis, data scientists require powerful back-end systems (data science platforms) to crunch raw data. Furthermore, the platforms have to provide an interactive mode of data analysis, which is required due to the iterative and inquisitive nature of performing data science.

3 Execution

We are planning to set up a center for data science as shown in Figure 2. We need two types of resources:

- *A. Bulut is with the Department of Computer Science, İstanbul Şehir University, İstanbul, Turkey, 34662. E-mail: ahmetbulut@sehir.edu.tr*
- *T. Arıcı is with the Department of Computer Science, İstanbul Şehir University, İstanbul, Turkey, 34662. E-mail: tarikarici@sehir.edu.tr*
- *O. Güzey is with the Department of Computer Science, İstanbul Şehir University, İstanbul, Turkey, 34662. E-mail: onurguzey@sehir.edu.tr*
- *B. Arslan is with the Department of Computer Science, İstanbul Şehir University, İstanbul, Turkey, 34662. E-mail: barisarslan@sehir.edu.tr*
- *A. Çakmak is with the Department of Computer Science, İstanbul Şehir University, İstanbul, Turkey, 34662. E-mail: alicakmak@sehir.edu.tr*

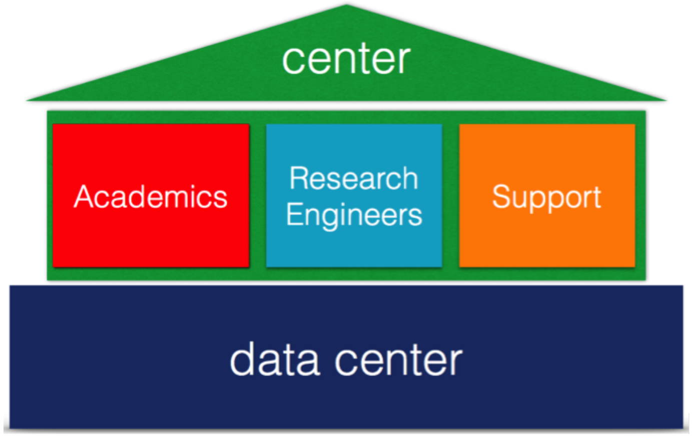


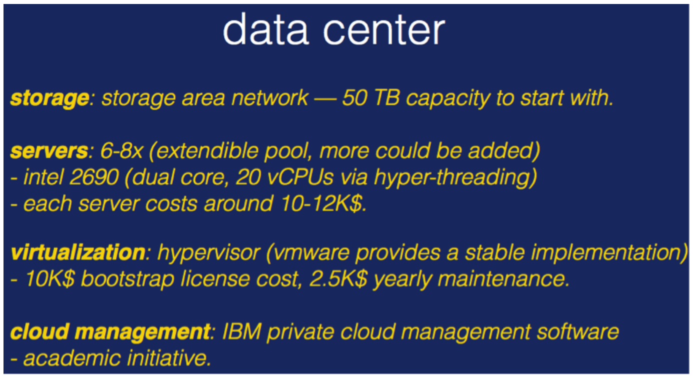
Fig. 2. The stakeholders and the building blocks of the center for data science.

human resources and hardware resources. Human resources refer to the academicians, the research engineers, and the support staff. All research engineers will be trained through our own Data Science masters program. The program will be executed with IBM's partnership. The partnership gives us access to a wide array of software, hardware, educational resources. We need to hire the support staff. The support staff consists of server admins, web admins, and project managers. The server admins will be responsible for operating the data center and provisioning the compute resources to the project teams. The web admins will be responsible for publishing and marketing the center activities, e.g., the ongoing projects, publications, grants, awards, news articles, workshops etc., on the web to the public. The project managers will be responsible for organizing the projects, managing the project funds and the related accounting.

Hardware resources refer to the physical compute capacity housed in the data center. The initial bootstrap configuration of the center is shown in Figure 3. The capex of this configuration is approximately $150,000$ USD in total. We expect that this infrastructure will grow over time with the increasing resource demand. Therefore, the infrastructure will initially be provided as an elastic private cloud. Different cloud offerings such as a hybrid or a purely public cloud will be considered when deemed necessary.

4 Masters Program

In order to fully grasp the opportunities present in the current environment awash with data, we designed a targeted graduate program for Data Science. Our institutional partnership with IBM will help us better this new graduate program going forward. Besides systems and tooling support, enhancing class experience with guest lectures from IBM personnel expert in the areas pertinent to course content will make sure that the practical implications and real problems to solve will always be considered up-front in the academia.



data center

storage: *storage area network — 50 TB capacity to start with.*

servers: *6-8x (extendible pool, more could be added)*
- intel 2690 (dual core, 20 vCPUs via hyper-threading)
- each server costs around 10-12K$.

virtualization: *hypervisor (vmware provides a stable implementation)*
- 10K$ bootstrap license cost, 2.5K$ yearly maintenance.

cloud management: *IBM private cloud management software*
- academic initiative.

Fig. 3. The initial bootstrap configuration of the data center.

Our masters program is offered with two options to choose from as:

Masters in Data Science with Thesis:

1) Students must complete at least 24 credits of course work (8 courses), in addition to a seminar course.
2) Students must complete a research thesis under the supervision of a faculty adviser.

Masters in Data Science without Thesis:

1) Students must complete at least 30 credits of course work (10 courses), in addition to a seminar course.
2) Students must complete and present a term project under the supervision of an adviser.

The program courses are shown in Table 1 where they are divided into three broad categories as theory, core, and application. The course descriptions for only a subset of the these program courses are listed below for reference:

- Data Engineering: Information retrieval concepts and data engineering skills on practical applications.
- Networks: Graph & Game theoretic analysis of Web, Social Networks, and Sponsored Search Markets.
- Data Visualization: Techniques to visualize high-dimensional data for insight discovery.
- Scalable Systems: How to build consumer facing Web systems and architectures that can scale.
- Big Data Analysis: Tools used for analyzing Big Data.
- Probabilistic Graphical Networks: How to establish relationships between entities and objects for probabilistic inference.
- Machine Learning: Theory behind well-established classification, regression, and clustering methodologies.
- Linear Dynamical Systems: Representation of dynamic systems in state space to understand their evolution over time.
- Optimization: Techniques used to optimize real world problems with real constraints.

We expect that all thesis students will complete core course requirements fully and build the necessary mathematical background via theory classes, while the non-

TABLE 1
The current curriculum of the Masters in Data Science program.

Theory	Core	Application
Elements of Statistical Learning	Data Engineering	Data Science for Business
Convex Optimization	Networks	Multimedia
Information Theory	Data Visualization	Service Based Startups
Linear Dynamical Systems	Big Data Analysis	Bioinformatics
Numerical Methods	Probabilistic Graphical Networks	
	Machine Learning	
	Scalable Systems	

TABLE 2
The suggested course plan for thesis and non-thesis options.

	Theory	Core	Application
w/ Thesis	1 or 2	4 or 5	2
w/o Thesis	0	6	4

thesis students will put more emphasis on the core courses and the application side of things. Therefore, our initial suggested course plans for thesis and non-thesis options are shown in Table 2, where the required number of classes to fullfill in order to graduate are allocated to three categories accordingly.

References

[1] John Gantz and David Reinsel, IDC iView ”Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East,” December 2012, sponsored by EMC.

[2] IBM InfoSphere, ”What is a data scientist”, http://www-01.ibm.com/software/data/infosphere/data-scientist/, retrieved on Aug 7, 2014.